Cities 96 (2020) 102410

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Multi-class twitter data categorization and geocoding with a novel computing framework

Sakib Mahmud Khan[a,*], Mashrur Chowdhury[a], Linh B. Ngo[b], Amy Apon[c]

[a] Glenn Department of Civil Engineering, Clemson University, Clemson, SC 29634, USA
[b] Computer Science Department, West Chester University, West Chester, PA 19383, USA
[c] School of Computing, Clemson University, SC 29634, USA

ARTICLE INFO

Keywords:
Social media
New York
Traffic operation
Short-term planning
Machine learning
Traffic management policy

ABSTRACT

This study details the progress in transportation data analysis with a novel computing framework in keeping with the continuous evolution of the computing technology. The computing framework combines the Labeled Latent Dirichlet Allocation (L-LDA)-incorporated Support Vector Machine (SVM) classifier with the supporting computing strategy on publicly available Twitter data in determining transportation-related events to provide reliable information to travelers. The analytical approach includes analyzing tweets using text classification and geocoding locations based on string similarity. A case study conducted for the New York City and its surrounding areas demonstrates the feasibility of the analytical approach. Approximately 700,010 tweets are analyzed to extract relevant transportation-related information for one week. The SVM classifier achieves > 85% accuracy in identifying transportation-related tweets from structured data. To further categorize the transportation-related tweets into sub-classes: incident, congestion, construction, special events, and other events, three supervised classifiers are used: L-LDA, SVM, and L-LDA incorporated SVM. Findings from this study demonstrate that the analytical framework, which uses the L-LDA incorporated SVM, can classify roadway transportation-related data from Twitter with over 98.3% accuracy, which is significantly higher than the accuracies achieved by standalone L-LDA and SVM.

1. Introduction

Traffic information is currently available through different private sources and navigation applications developed by private companies, such as Waze, Google, or Apple. At the same time, public agencies, specifically law enforcement agencies, must collect, validate, and disseminate incident information, as they are primarily responsible for traffic management and safety. A 2015 survey found that most state transportation agencies collect traffic data from sensors and through third parties, such as INRIX, and then use web sites and Dynamic Message Signs to disseminate traffic information to travelers (Fries et al., 2015). In the study conducted by Fries et al. (2015), based on the survey responses, researchers emphasized the need for improvement in methods and technologies for travel time data collection. As stated in a USDOT (2018) report, transportation applications using real-time data increases the operational and safety benefits by generating data helpful for making informed travel decisions (USDOT, 2018). Given the importance of the quality and availability of traffic data for providing reliable transportation services, tools that provide accurate, timely and accessible data to support traffic management and planning practices related to roadway traffic information collection and dissemination are essential. In addition to navigation applications developed by private companies, social media platforms like Twitter produce publicly available data that can provide 'where', 'what' and 'when' information about any traffic incident event. For example, "Incident on #MontaukBranch EB at Jamaica Station" tweet says where (i.e., at MontaukBranch EB, Jamaica Station) and what event (i.e., incident) happened. Another example tweet, "real confused as to why the workers aren't out here cleaning the roads!!" tells what event (i.e., there are obstructions or debris on the road), but the tweet itself does not tell where the event happened unless tweet has geolocation information available beyond the tweet text. In both tweet examples, the time of tweet generation is provided by Twitter. While Twitter has been analyzed as a potential source of traffic data (D'Andrea, Ducange, Lazzerini, & Marcelloni, 2015; Gu, Qian, & Chen, 2016), tweets do not always have geolocation information available. Also, since drivers should not tweet while driving, Twitter data is most appropriate as support for traffic incident-related data in which the tweets from the general public

* Corresponding author.
E-mail addresses: sakibk@g.clemson.edu (S.M. Khan), mac@clemson.edu (M. Chowdhury), lngo@wcupa.edu (L.B. Ngo), aapon@clemson.edu (A. Apon).

https://doi.org/10.1016/j.cities.2019.102410
Received 9 October 2018; Received in revised form 8 July 2019; Accepted 13 July 2019

originate from stopped vehicles or the passengers within (Pratt, Morris, Zhou, Khan, & Chowdhury, 2019).

In this paper, the term 'tweet' refers to the message or status update from a Twitter user account, which cannot exceed the 140 character limit (the size of tweets has been extended to 280 characters since the time of this study). Although Twitter provides data generated by numerous users from a specific region, analyzing the raw streaming data in real-time and providing useful feedback based on the analysis are challenging. The research objective is to develop a parallel-computing based analytical framework to accurately categorize and reliably geocode tweets for the transportation-related events. This contribution of this paper entails developing and evaluating: (a) the Labeled Latent Dirichlet Allocation (L-LDA)-incorporated Support Vector Machine (SVM) classifier to classify tweets with supporting distributed computing framework to support roadway transportation operations and (b) the string-similarity based location identification system.

After analyzing the collected tweets from a specific region using the Natural Language Processing (NLP) techniques, transportation-related tweets are extracted with SVM, a supervised classification technique. SVM is used to identify transportation-related tweets from the whole Twitter dataset for each day, and the Clemson University Palmetto supercomputing cluster is used to support parallel computations to develop SVM models. The motivation of using this parallel computation framework, to classify almost 700,010 tweets in this study, is to minimize the computation time for the SVM training phase compared to single node-based computation. After identification, the transportation-related tweets are classified via three supervised classification techniques: L-LDA, SVM, and L-LDA incorporated SVM. L-LDA is a supervised credit attribution method, whereas L-LDA and L-LDA incorporated SVM have not been used to identify transportation-related events in earlier research. It has been previously determined that L-LDA performs as well as or better than SVM for multi-label text classification (Ramage, Hall, Nallapati, & Manning, 2009). The motivation for integrating L-LDA with SVM in this study is to improve the performance of SVM in classifying tweets. In the L-LDA incorporated SVM technique, topic distribution probability for each tweet generated by L-LDA is used by SVM classifier to categorize the tweets in multiple classes (i.e., incident, congestion, special event, construction, and other events). Accuracies of SVM, L-LDA, and L-LDA incorporated SVM classifiers are measured with respect to the labels manually assigned to the tweets.

According to Title 23 of the Code of Federal Regulations, real-time highway information programs, including statewide incident reporting system, must be 85% accurate as a minimum (GPO, 2011). It can be inferred, from this code, that it is possible to use Twitter as a potential standalone tool to compile and classify roadway transportation events if the accuracy is above the 85% threshold. Following the text classification, the tweets are geocoded. Using the analytical framework presented in this study, a case study is conducted for New York City (NYC) and its surrounding areas. The following sections discuss the previous studies related to twitter data analysis, analytical framework for this study, and a case study using the analytical framework.

2. Literature review

Twitter data are used for assessing various events (D'Andrea et al., 2015; Gu et al., 2016; He, Boas, Mol, & Lu, 2017; Purohit et al., 2014; Qian, 2016; Roberts et al., 2018; Tang et al., 2017) including natural disasters, mass emergency, acts of terrorism, extreme weather events, political protests, and transportation events. In a study conducted by Mirończuk and Protasiewicz (2018), the authors have reviewed recent research to understand the general approach of text classification practices and identify the future research questions related to text classification (Mirończuk & Protasiewicz, 2018). The most common research for text classification includes the use of supervised learning methods and involves a number of steps including data acquisition, data labeling, feature construction, feature weighing, feature selection, classification model training, and assessment. The authors have identified overfitting of the text classification models, dynamic classifier selection, multi-lingual text analysis, text stream analysis, sentiment analysis and ensemble-learning methods as the emerging research topics in text classification.

2.1. Tweet classification with machine learning

Once Twitter data are collected, their contents are analyzed. This is a difficult process, as Twitter data is often characterized as "vast, noisy, distributed, unstructured, and dynamic" (Gundecha & Liu, 2014). Therefore, machine-learning techniques are integral to the process of mining content for decision-making purposes. These machine learning techniques are categorized into three primary areas, supervised (Kotsiantis, 2007), semi-supervised (Zhu, 2006), and unsupervised (Hastie, Friedman, & Tibshirani, 2001). A supervised learning algorithm uses training data with known outcomes. The learning algorithm can gradually adjust its parameters to generate results from training data so that these results match most closely with the known outcomes. For unsupervised learning, there are no known outcomes, and the algorithm will attempt to extract the pattern from the data itself. Semi-supervised learning techniques contain a mixture of both by using a small set of training data with known outcomes and a majority of training data without known outcomes. The evidence of the various degrees of success in applying different machine learning techniques to analyze social media contents is well known (D'Andrea et al., 2015; Ramage et al., 2009). For this specific study, a supervised machine learning technique, SVM (Cortes & Vapnik, 1995), is selected to automate the process of identifying transportation/non-transportation tweets, and L-LDA (Ramage et al., 2009), another supervised technique, is selected to model the topics of the classified tweets. SVM facilitates the utilization of kernel functions to develop hyperplane(s) within the feature space of the observation to classify the observations into different distinctive groups. Supervised LDA (s-LDA) methods are used to identify the label of the tweets by simply restricting the topic model to use only the topics corresponding to the training dataset's label set. A s-LDA classifier is used by Gu et al. (2016), where the authors found that 51% of the geo-codable tweets can be classified with their s-LDA classifier.

2.2. Twitter data for transportation applications

In earlier investigations of the reliability and accuracy of social media data for unplanned transportation events (i.e., incidents, congestion), various methods (i.e., machine learning, statistical analysis) were proposed to extract necessary data from user-focused contextual information that is shared in the social media platform. To determine real-time incident information, Twitter data were analyzed using machine learning technique that incorporated semantic web technology (i.e., Linked Open Data Cloud) and features from tweets and LOD data for tweet classification (i.e., car crash class, shooting class and fire class) (Schulz & Ristoski, 2013). The proposed model achieved about 89% accuracy for classifying tweets. They concluded that even with very few social media posts, this method is capable of detecting incidents. For traffic congestion monitoring, (Chen, Chen, & Qian, 2014) developed a statistical framework that integrated both Hinge-loss Markov Random Fields and a language model. Evaluations were performed over different spatial-temporal and other performance metrics on the collected tweet and INRIX probe datasets. The two major U.S. cities used in this study were Washington D.C. and Philadelphia, PA. Based on their analysis, (Chen et al., 2014) found that Twitter data can supplement traditional road sensor data to assess traffic operational conditions. The authors from (Sakaki, Matsuo, Yanagihara, Chandrasiri, & Nawa, 2012) study created a system to distribute important event-related information to vehicle drivers, including the location information and temporal information. Tweets were classified as either traffic